GeoX-Bench: Benchmarking Cross-View Geo-Localization and Pose Estimation Capabilities of Large Multimodal Models

Yushuo Zheng[1,2], **Jiangyong Ying**[3], **Huiyu Duan**[1*], **Chunyi Li**[1,2], **Zicheng Zhang**[1,2], **Jing Liu**[4], **Xiaohong Liu**[1,5*], **Guangtao Zhai**[1,2*]

[1]Shanghai Jiao Tong University
[2]Shanghai Artificial Intelligence Laboratory
[3]China Telecom
[4]Tianjin University
[5]Shanghai Jiao Tong University Sichuan Research Institute
{yushuozheng, huiyuduan, lcysyzxdxc, zzc1998, xiaohongliu, guangtaozhai}@sjtu.edu.cn,
jiangyingyung@chinatelecom.cn, jliu_tju@tju.edu.cn

Abstract

Large multimodal models (LMMs) have demonstrated remarkable capabilities across a wide range of tasks, however their knowledge and abilities in the cross-view geo-localization and pose estimation domains remain unexplored, despite potential benefits for navigation, autonomous driving, outdoor robotics, *etc*. To bridge this gap, we introduce **GeoX-Bench**, a comprehensive Benchmark designed to explore and evaluate the capabilities of LMMs in cross-view Geo-localization and pose estimation. Specifically, GeoX-Bench contains 10,859 panoramic-satellite image pairs spanning 128 cities in 49 countries, along with corresponding 755,976 question-answering (QA) pairs. Among these, 42,900 QA pairs are designated for benchmarking, while the remaining are intended to enhance the capabilities of LMMs. Based on GeoX-Bench, we evaluate the capabilities of 25 state-of-the-art LMMs on cross-view geo-localization and pose estimation tasks, and further explore the empowered capabilities of instruction-tuning. Our benchmark demonstrate that while current LMMs achieve impressive performance in geo-localization tasks, their effectiveness declines significantly on the more complex pose estimation tasks, highlighting a critical area for future improvement, and instruction-tuning LMMs on the training data of GeoX-Bench can significantly improve the cross-view geo-sense abilities.

Code — https://github.com/IntMeGroup/GeoX-Bench

1 Introduction

Accurately determining the geographic location and pose of the camera is a fundamental challenge in computer vision and robotics, with broad implications for applications such as autonomous driving, robotic navigation, and embodied AI agents operating in real-world environments. These scenarios often involve degraded or unavailable GPS signals and magnetic interference, necessitating robust visual-based geospatial reasoning, *i.e.*, cross-view geo-localization and pose estimation. However, the inherent complexity of the physical world, manifested in its large-scale spatial layout and diverse visual appearance, makes this task particularly difficult. Recent advances in large multimodal models (LMMs), such as GPT-4V (OpenAITeam 2024) and LLaVA (Liu et al. 2023a) have demonstrated impressive capabilities in free-form visual reasoning across a range of tasks. Despite this progress, their proficiency in geospatial grounding, the ability to understand and reason about geographic location and spatial orientation, remains largely unexamined, due to the absence of standardized, large-scale benchmarks specifically designed to assess such capabilities.

Existing efforts to evaluate the geographic understanding of LMMs have primarily focused on satellite image geo-localization, which provides only coarse-grained location information and lacks the estimation of camera pose (Hu et al. 2018). Another task, geo-guessing, requires models to predict the approximate location of a ground-level image only based on its visual content (Weyand, Kostrikov, and Philbin 2016). However, this task does not involve the cross-view geo-localization and pose estimation problem, which aims to determine both the geographic location and pose based on the captured ground-level image and the aerial perspective image. This more comprehensive task demands a deeper and more structured understanding of spatial relationships in the visual scene, presenting a more rigorous test of LMMs' geospatial reasoning capabilities.

In this paper, we introduce a new benchmark termed **GeoX-Bench** for evaluating the capabilities of LMMs on cross-view geo-localization and pose estimation tasks.GeoX-Bench comprises 10,859 panoramic-satellite image pairs covering 128 cities across 49 countries, along with 755,976 carefully curated corresponding question-answering (QA) pairs. Among these, 42,900 QA pairs are curated for standardized benchmarking, while the remaining are intended to enhance model capabilities through instruction tuning. GeoX-Bench defines two core tasks, including **(1) geo-localization**, *i.e.*, predicting the geographic location of a ground-level perspective image given a satellite image; and **(2) geo-pose estimation**, *i.e.*, inferring the camera's orientation of a ground-level perspective image given a satellite image.

Based on GeoX-Bench, we establish rigorous evaluation protocols, and evaluate the capabilities of 25 state-of-the-

*Corresponding author.

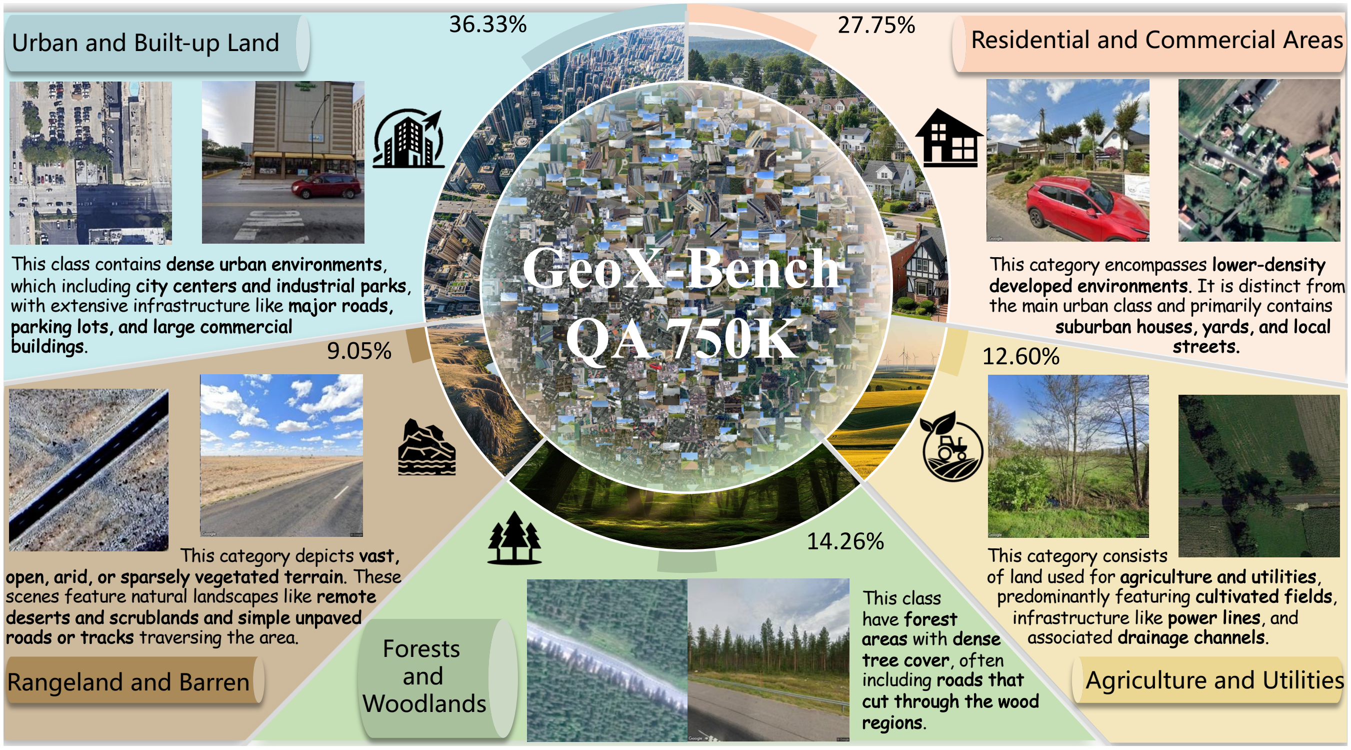

Figure 1: Geographic composition of the **GeoX-Bench** dataset by land cover type. The benchmark is weighted toward developed regions, with Urban/Built-up (36.33%) and Residential/Commercial (27.75%) areas constituting the majority. Natural and rural environments, including Forests (14.26%), Agriculture (12.60%), and Rangeland/Barren (9.05%), provide a geospatially representative representative diverse settings for evaluation.

art LMMs and instruction-tuned LMMs on cross-view geo-localization and pose estimation tasks. Our experimental results reveal that while contemporary LMMs achieve promising performance on geo-localization, their capabilities degrade considerably on the more challenging pose estimation task. Moreover, the significant improvement brought by the instruaction-tuning suggests that the capabilities of LMMs on both cross-view geo-localization and pose estimation tasks can be further improved.

Our main contributions are as follows:

- We introduce GeoX-Bench, a large-scale benchmark comprising 10,859 panoramic-satellite image pairs and 755,976 curated QA pairs designed for evaluating LMMs' capabilities on cross-view geo-localization and pose estimation.
- We propose rigorous evaluation metrics for both tasks and benchmark 25 state-of-the-art LMMs and instruction-tuned variants.
- We provide a comprehensive analysis of the performance of LMMs on both tasks, highlighting the strengths and weaknesses of current models and identifying critical areas for future improvement.

2 Related Work

2.1 Large Multimodal Models and Their Spatial Reasoning Capability

The rapid evolution from Large Language Models (LLMs) (Brown et al. 2020) to Large Multimodal Models (LMMs) such as LLaVA (Liu et al. 2023a), GPT-4 (OpenAITeam 2024), and Gemini (GeminiTeam 2024) has unlocked advanced capabilities in processing both text and vision (Duan et al. 2025, 2024; Yang et al. 2025b,a). LMMs have been increasingly applied to embodied intelligence (Mai et al. 2023), where spatial awareness is essential (Jin and Jia 2025). Early explorations, such as LLMGeo (Wang et al. 2024b) for coarse location reasoning, LLaVA-3D (Zhu et al. 2024) for pose understanding, and All-Angles Bench (Yeh et al. 2025), demonstrate emerging capabilities, while recent competitive benchmarks (Zheng et al. 2025) broaden evaluation perspectives. However, existing studies address spatial skills in isolation and lack the precision required for real-world applications. To address this, our work proposes fine-grained cross-view localization and pose-estimation tasks that more directly evaluate integrated spatial reasoning. Additionally, the construction and organization of evaluation tasks play a critical role, since well-structured benchmarks that span diverse locations and viewpoints can more reliably expose real-world spatial limitations beyond isolated skills.

Figure 2: Illustration of the **GeoX-Bench** benchmark tasks. The tasks include heading estimation with known or unknown camera locations, location verification on a satellite map, location selection, and map selection from candidates. These tasks evaluate models' abilities to reason over ground-to-satellite image pairs for localization and pose understanding.

2.2 Cross-View Geo-Localization and Geo-Pose Estimation

Evaluation methods have evolved alongside VLMs. Foundational benchmarks such as VQA (Goyal et al. 2016) established multimodal reasoning, while recent works like MME (Fu et al. 2024) and MMMU (Yue et al. 2024) assess complex, expert-level tasks. Broader evaluation platforms (Zhang et al. 2025) and perceptual studies (Chen et al. 2025) further diversify assessment frameworks. In remote sensing, benchmarks like RSVQA (Lobry et al. 2020) and GeoChat (Kuckreja et al. 2024) focus primarily on aerial imagery, while immersive-environment benchmarks (Ji et al. 2025a,c,b, 2024) assess medical or panoramic understanding. Despite these efforts, a gap persists: systematic evaluation of ground-to-satellite spatial reasoning. Existing cross-view datasets—CVUSA (Workman, Souvenir, and Jacobs 2015), VIGOR (Zhu, Yang, and Chen 2021), OmniCity (Li et al. 2023), LLMGeo (Wang et al. 2024b), CVGlobal (Zhu and Yang 2024)—provide strong foundations yet remain fragmented and geographically biased. **GeoX-Bench** unifies six diverse datasets to create the first geographically broad, multi-view benchmark designed specifically to assess both fine-grained geo-localization and orientation reasoning, as illustrated in Figure 1.

3 GeoX-Bench

3.1 Overview

We propose **GeoX-Bench**, a novel benchmark specifically designed for evaluating the capability of Large Multimodal Models (LMMs) in understanding geo-localization and pose estimation in outdoor environments. **GeoX-Bench** comprises a training set of 10,684 panorama-satellite pairs and a test set of 175 pairs, distributed across the world as shown in Figure 1. This corresponds to a total of over 750,000 question-answer pairs, systematically divided into training and test sets while maintaining a consistent question and answer distribution. For detailed statistical information, please refer to the supplemental material. GeoX-Bench is characterized by: (1) comprehensive cross-view and multimodal information, where each image pair or group includes at least one ground-level image and one satellite image; and (2) extensive variability in land types, facilitating the assessment of LMM performance across diverse scenarios. GeoX-Bench encompasses seven specific tasks, covering both **geo-localization** and **geo-pose estimation** as shown in Figure 2.

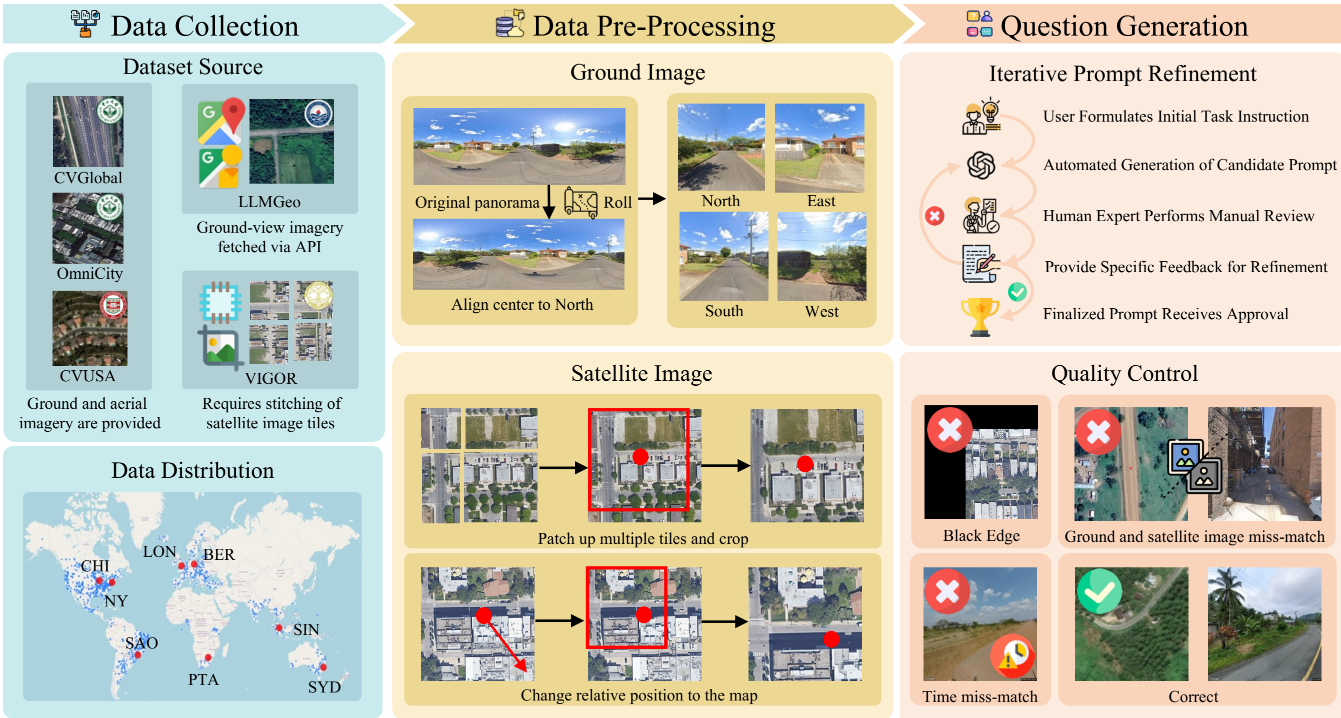

Figure 3: The **GeoX-Bench** data curation pipeline, from source sampling to final quality control. We first sample ground-satellite pairs from four existing datasets to ensure broad geographic coverage. In the pre-processing stage, ground-level panoramas are programmatically rotated to align to a consistent North orientation before cardinal views are extracted, while corresponding satellite imagery is stitched and cropped. An iterative, LLM-assisted framework with human oversight is used for question prompt generation. A final quality control stage removes data with visual artifacts or cross-view inconsistencies, such as spatial or temporal mismatches, to ensure benchmark integrity.

Comparison with Existing Benchmarks While general benchmarks such as MMBench (Liu et al. 2023b) effectively assess the overall performance of LMMs, they predominantly evaluate single-view capabilities. In contrast, **GeoX-Bench** uniquely enables the evaluation of LMMs across different viewpoints and scenarios. Benchmarks like LLMGeo (Wang et al. 2024b) assess model capabilities in geographical location inference but are limited to single-view inputs, lacking precise geo-localization accuracy and neglecting pose estimation capabilities. Additionally, benchmarks such as All-Angles Bench (Yeh et al. 2025), though designed to evaluate multi-view understanding, differ fundamentally from GeoX-Bench by providing multiple views from a single modality. Conversely, GeoX-Bench distinguishes itself by incorporating both ground-level and satellite imagery, thereby enabling robust multi-modal and cross-view evaluation.

3.2 Benchmark Tasks

Pose Estimation Given a ground-level image and a north-aligned satellite map, the model must infer the camera's absolute heading (North, East, South, or West). In the *Fixed Position* variant, the camera is placed at the map's center; in the *Random Position* variant, the camera is placed at an arbitrary location within the map, requiring the model to reason about orientation without positional cues.

Localization The model must determine whether the ground scene is contained within the given satellite map. The *With Location Prior* variant assumes the camera is fixed at the map's geometric center, while the *Without Prior* variant allows the camera to be anywhere inside the map, requiring broader spatial reasoning.

Intra-Map Localization Assuming the ground image is captured from somewhere within the map, the model must predict the camera's relative position (e.g., top, bottom, left, right) with respect to the map center.

Cross-Map Retrieval Given a ground photograph and four candidate satellite maps, the model must identify the correct matching map. The *Standard* variant places the camera at the center of the correct map, while the *Random* variant places it at an arbitrary position, forcing the model to rely on holistic scene understanding.

3.3 Benchmark Curation

Data Collection As illustrated in Figure 3, the GeoX-Bench dataset was created by sampling data from several established datasets. To ascertain model robustness, we sampled a balanced number of data points from the CVGlobal, CVUSA, OmniCity, LLMGeo, and VIGOR datasets. Since temporal synchronization between satellite and ground-level

Type	Model	Pose Estimation		Localization		Intra & Cross map Selection			Average
		Fixed	Random	w/ Prior	w/o Prior	IM Rnd.	CM Std.	CM Rnd.	
Random	Random Selection	25.00	25.00	50.00	50.00	11.11	25.00	25.00	30.16
Open Source	DeepSeek-VL-7B	25.36	25.07	50.64	51.48	11.63	19.91	22.51	29.51
	InternVL2-2B	24.00	24.38	50.09	49.79	10.02	25.73	25.74	29.96
	InternVL2-4B	24.45	24.45	50.18	50.31	11.92	24.18	24.12	29.95
	InternVL2-8B	26.27	24.86	62.09	60.73	11.30	30.64	28.33	34.89
	InternVL2-40B	25.09	25.04	70.09	66.63	11.33	29.91	28.75	36.69
	InternVL2-76B	25.27	25.21	63.18	62.43	10.22	34.91	33.61	36.41
	InternVL3-2B	25.73	25.75	57.09	58.48	11.41	28.27	27.03	33.40
	InternVL3-8B	25.36	25.12	68.36	65.41	11.47	43.64	41.21	40.08
	InternVL3-38B	27.64	25.09	76.27	71.86	6.67	69.73	64.82	48.87
	InternVL3-78B	27.64	27.89	74.64	70.24	13.29	68.45	61.45	49.09
	LLaVA-Interleave-7B	25.73	24.79	51.27	52.09	11.41	24.00	24.46	30.54
	LLaVA-One-Vision-7B	22.82	24.51	58.91	56.76	11.42	26.64	27.81	32.69
	mPLUG-Owl3-7B	24.00	18.85	53.82	53.16	10.41	19.00	19.03	28.32
	MS-Phi3d5	24.18	24.21	54.27	52.27	11.07	21.82	22.93	30.11
	Qwen2.5VL-3B	27.73	25.54	50.00	49.81	11.46	24.73	24.89	30.59
	Qwen2.5VL-7B	23.27	24.99	65.00	63.10	11.26	44.00	38.30	38.56
	Qwen2.5VL-32B	22.27	19.91	72.27	70.87	11.69	56.82	52.53	43.76
	Qwen2.5VL-72B	29.91	27.53	71.73	68.69	12.73	65.55	61.98	48.30
	Qwen2VL-2B	24.91	24.84	47.91	50.42	11.18	24.91	27.11	30.18
	Qwen2VL-7B	25.00	24.16	64.73	61.15	11.99	32.09	33.07	36.03
	Qwen2VL-72B	27.55	26.71	71.18	66.79	11.76	70.27	60.28	47.79
Closed Source	Claude-Sonnet-4	29.55	37.12	68.18	62.88	14.39	43.94	56.82	44.70
	Gemini-2.5-Pro	45.45	41.67	87.88	79.55	18.94	85.61	77.27	62.34
	GPT-4o	33.33	26.52	90.15	82.58	15.15	80.30	81.06	58.44
	o3	50.00	**46.97**	84.09	81.06	18.94	84.09	78.79	63.42
Instruaction Tuning (LoRA)	InternVL2-8B	25.00	24.87	**98.82**	84.82	24.58	87.45	78.73	60.61
	InternVL3-8B	**55.36**	40.80	98.45	**91.97**	**36.33**	**91.73**	**84.80**	**71.35**
	Qwen2.5VL-7B	51.36	36.87	98.09	81.08	35.80	88.55	79.94	67.38
	Qwen2VL-7B	53.27	38.73	98.27	82.85	35.62	90.27	83.71	68.96

Table 1: Performance comparison of various vision-language models across different tasks, organized by model type. The highest score in each column is shown in **bold**, and the second-highest is underlined. The IM Rnd. means Intra-Map Localization with random location, CM Std. means Cross-Map Retrieval with location at center of the map prior and CM Rnd. means Cross-Map Retrieval without location prior.

imagery is crucial, we adopted methodologies from prior research. Specifically, we re-accessed the Google Maps API to obtain current satellite imagery corresponding to the street-view images in the datasets, ensuring both views were captured between 2024 and 2025.

Data Pre-processing For ground-level images, panoramic views from datasets like CVGlobal and OmniCity were first rotated to a standard North-facing orientation. Following prior methodology (Wang et al. 2024b), each panorama was then transformed into four images with a 90-degree Field of View (FoV), each oriented towards a cardinal direction (North, East, South, and West). For aerial images, particularly from the VIGOR dataset where satellite and ground views do not have a one-to-one correspondence, we merged and resized satellite imagery to a uniform 512×512 pixels. This process ensures the ground-level perspective is precisely centered within its corresponding satellite image. For map randomization tasks, we used off-center cropping to reposition these central locations arbitrarily.

Prompt Generation We used GPT-4o for prompt generation. Each prompt consists of a system prompt and a user prompt. The system prompt describes the input data, defines the task, and concludes with explicit formatting instructions and examples. The user prompt provides the relevant images, reiterates the task, and reinforces the required output format. After generation, all prompts were manually reviewed to ensure quality and adherence to requirements.

Quality Control Samples exhibiting visual artifacts—such as black padding, corrupted tiles, or missing imagery—were immediately removed. The remaining pairs were then scrutinized for cross-view and temporal coherence. Each street-level photograph had to spatially align with its satellite coun-

Model	H_{norm}	D_{KL}	p_{mode}	a_{mode}
InternVL2-4B	0.4681	0.9337	0.6915	0.2982
InternVL2-8B	0.5587	0.9021	0.6975	0.3063
InternVL2-40B	0.7145	0.6098	0.5786	0.3005
InternVL2-76B	0.8029	0.4650	0.4951	0.3000
InternVL3-2B	0.5228	0.9163	0.7197	0.3031
InternVL3-8B	0.7620	0.4981	0.5269	0.3095
InternVL3-38B	0.7776	0.5160	0.5412	0.3009
InternVL3-78B	0.8061	0.3761	0.4961	0.3112
Qwen2.5VL-3B	0.5053	0.8679	0.7063	0.3000
Qwen2.5VL-7B	0.7020	0.6326	0.5955	0.3067
Qwen2.5VL-32B	0.8312	0.4417	0.4579	0.2973
Qwen2.5VL-72B	0.8631	0.3412	0.4359	0.3138
Qwen2VL-2B	0.4412	1.0985	0.7342	0.3012
Qwen2VL-7B	0.7325	0.5981	0.5574	0.3051
Qwen2VL-72B	0.8216	0.4126	0.4520	0.3108

Table 2: Average option-bias statistics across several tasks, including Heading Estimation and Map Selection, for various LMMs. We report the normalized entropy (H_{norm}), KL divergence from uniform (D_{KL}), the frequency of the most-predicted answer (p_{mode}), and the accuracy achievable by only guessing that single answer (a_{mode}).

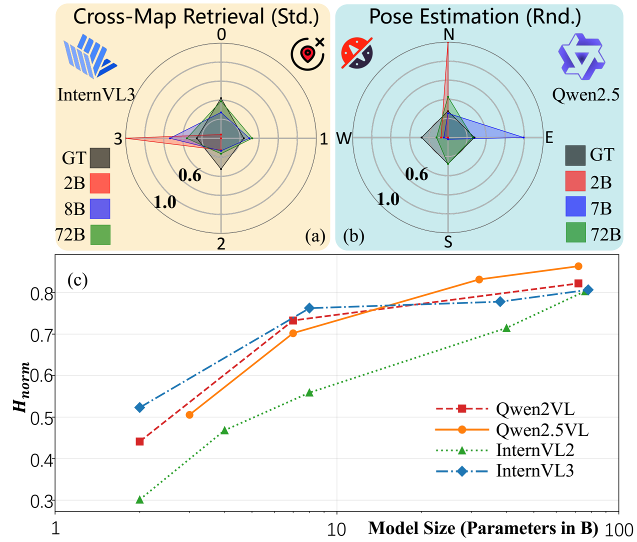

Figure 4: Model scale reduces choice bias across tasks and architectures. Smaller models show strong option preference in both *Cross-Map Retrieval* and *Pose Estimation*, while larger models produce more uniform predictions. Normalized entropy (H_{norm}) increases with parameter count, indicating improved calibration and reduced reliance on choice priors.

terpart, and the capture dates had to be close enough to prevent significant appearance changes. Any instances with clear mismatches (e.g., an urban alley paired with a rural tile) or large time gaps were discarded.

4 Experiments

4.1 Evaluation Setup

Our analysis includes 21 open-source models and 4 leading closed-source models to provide a comprehensive comparison. The open-source models tested cover several prominent families, including the InternVL series (Chen et al. 2024; Zhu et al. 2025), the QwenVL series (Wang et al. 2024a; Bai et al. 2025), and models from the LLaVA family (Li et al. 2024b,a). We also include other notable models such as mPLUG-Owl3-7B (Ye et al. 2024) and Phi-3.5-Vision-Instruct (Abdin et al. 2024). The close-source models including Claude-Sonnet-4 (Team 2024), Gemini-2.5-Pro (Comanici et al. 2025), GPT-4o (OpenAITeam 2024) and o3 (OpenAI 2025).For additional context, we include benchmarks of four instruction turned model on a portion of the GeoX-Bench training dataset. To ensure reproducibility, we set the temperature to 0 and perform greedy decoding.

To investigate the benefits of task-specific training, we conduct instruction turning on representative models from the Qwen-VL and InternVL families. Our finetuning dataset is a curated subset of the GeoX-Bench training data, containing 31,392 question-answer pairs derived from 654 unique locations. The training is focused exclusively on the three non-randomized tasks: Pose Estimation (Fixed), Localization with Prior and Cross-Map Selection. And the Intra-Map Localization task.

We employ the parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al. 2022) technique for all finetuning experiments. The models are trained for a total of 4 epochs on NVIDIA A800 GPU, utilizing the SWIFT (Zhao et al. 2024) framework. During training, the vision transformer (ViT) backbone remains frozen to preserve its learned general-purpose visual features. We adopt a bfloat16 mixed-precision setup and leverage Flash Attention (Dao 2024) to optimize computational efficiency.

Key hyperparameters are kept consistent across all instruction turning runs: a learning rate of 1×10^{-4}, a per-device batch size of 2, and a warmup ratio of 0.05. For LoRA, we configure the rank (r) to 8, the scaling factor (α) to 32, and apply the adaptation to all linear layers. The total training time is approximately 20 hours for QwenVL models and 30 hours for InternVL models. For final evaluation, we select the checkpoint that achieves the highest word accuracy on the validation set that split from the instruction turning data.

4.2 Main Results

LMM Performance Trends The evluation data across all 7 tasks are shown in Table 1, which reveals three key insight as following.(**a**)Task difficulty: Geo-localization is significantly easier than pose estimation. Even 2B parameter models outperform random chance on localization tasks, whereas no model surpasses 30% accuracy on heading estimation without fine-tuning.(**b**)Scaling laws: Accuracy generally increases with model scale up to around 32B parameters, after which performance tends to saturate for pre-trained models.(**c**)Fine-tuning impact: Instruction turning drastically improves performance. Notably, the **InternVL3-8B** model, when fine-tuned on our training set, outperforms all other models, including the much larger 72B parameter variants, demonstrating the high value of task-specific data.

Option Preference of LMMs To analyze how Large Language Models (LMMs) distribute their choices, we employ several metrics. First, to quantify the dispersion of selections across a fixed set of choices, we define the empirical probability p_i for each of the k possible outcomes based on the model's raw selection counts n_i:

$$p_i = \frac{n_i}{\sum_{j=1}^{k} n_j}$$

To measure the uniformity of this choice distribution, we employ **normalized entropy**. The Shannon entropy H of this discrete distribution is given by $H = -\sum_{i=1}^{k} p_i \log_2 p_i$. To make this metric comparable across tasks with a different number of choices k, we normalize H by the maximum possible entropy, $\log_2 k$. The resulting metric, H_{norm}, ranges from 0 (complete concentration on a single choice) to 1 (a perfectly uniform distribution):

$$H_{\text{norm}} = \frac{H}{\log_2 k} = -\frac{1}{\log_2 k} \sum_{i=1}^{k} p_i \log_2 p_i$$

In our experiments, we observe a clear trend: as a model's size increases, its normalized entropy H_{norm} also increases. This indicates that *larger models tend to distribute their selections more uniformly*. As shown in Figure 4, the positive correlation underscores that bigger models exhibit higher normalized entropy and thus less extreme preference for any single choice.

To directly measure the strength of a model's preference for its single favorite option, we define the **mode probability**, p_{mode}, as the empirical probability of the most frequently selected choice, $k^\star$:

$$p_{\text{mode}} = \max_k p_k$$

A high p_{mode} signifies a strong concentration on one choice. We can evaluate the impact of this primary bias by calculating the **mode-only baseline** accuracy, a_{mode}, which is the accuracy achieved if the model only ever predicted this single choice $k^\star$:

$$a_{\text{mode}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[y_i = k^\star]$$

where y_i is the ground-truth answer. Finally, to quantify the overall deviation of a model's answer distribution $\mathbf{p}$ from a uniform random guess $\mathbf{u}$, we use the **bias magnitude** as captured by the Kullback-Leibler (KL) divergence:

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{u}) = \sum_{i=1}^{K} p_i \log_2 \frac{p_i}{1/K}$$

Empirical Findings The data in Table 2 reveals a clear pattern: **smaller models exhibit a strong choice bias that diminishes as model scale increases**. Models in the 2B to 4B parameter range consistently display a pronounced preference for a single option. For example, Qwen2VL-2B assigns over 73% of its predictions to its most-favored choice ($p_{\text{mode}} = 0.7342$), resulting in a high KL divergence from uniform ($D_{\text{KL}} = 1.0985$ bits) and a low normalized entropy ($H_{\text{norm}} = 0.4412$).

Notably, this predictive bias is a consistent in a model. The accuracy of a **mode-only baseline** (a_{mode}), which exclusively uses the model's favorites answer, consistently lands around **0.30**. This performance is very close to the average accuracy result shows in Table 1. This suggests that smaller models rely on a simple way which is always give the preferred answer rather than understanding complex task semantics.

As model size increases, this dependence on a single choice lessens. For instance, the 72B parameter models show a much more uniform choice distribution, with p_{mode} values dropping below 0.45 and normalized entropy H_{norm} rising above 0.82. These findings underscore that the apparent choice bias in smaller models is a productive heuristic, allowing them to outperform random chance by exploiting a systematic prior. Acknowledging this artifact is crucial for correctly interpreting the capabilities of low-capacity models

5 Conclusion

In this work, we introduced **GeoX-Bench**, a large-scale benchmark for evaluating cross-view geo-localization and pose estimation in large multimodal models (LMMs). With over 10,000 aligned satellite–ground pairs and 750k QA across seven tasks, GeoX-Bench exposes spatial reasoning challenges that prior benchmarks cannot capture.

Our evaluation of more than 25 LMMs reveals a clear gap between coarse geo-localization and fine-grained pose estimation. While many high-parameter models exceed 70% accuracy on localization tasks, they consistently fail to surpass 30% on pose estimation as well as intra-map localization which require instance level recognition for both satellite image and the ground image which is a gap not solved by simply scaling parameters.This limitation persists even after general-purpose instruction tuning, suggesting a core architectural weakness. Although geometric instruction tuning offers notable improvements, it does not fully bridge the gap. Furthermore, we find the performance of smaller models is often artificially inflated by hard-coded biases, confirming that robust, cross-view geometric reasoning remains a significant, unsolved challenge for current LMMs.

Overall, our findings show that current LMMs can perform coarse localization but struggle with precise orientation and cross-view geometric reasoning. **GeoX-Bench** thus serves as a critical benchmark for driving future progress in geometric intelligence, particularly for safety-critical applications such as autonomous navigation and embodied AI.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 62225112, 62301310, 62401365, 62271312, 62132006, and U24A20220), the Sichuan Science and Technology Program (Grant 2024NSFSC1426), and the China Postdoctoral Science Foundation (Grants BX20250411 and 2025M773473).

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; and et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Conference on Neural Information Processing Systems*.

Chen, Z.; Tian, Y.; Sun, Y.; Sun, W.; Zhang, Z.; Lin, W.; Zhai, G.; and Zhang, W. 2025. Just Noticeable Difference for Large Multimodal Models.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; and et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Dao, T. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *Proceedings of International Conference on Learning Representations*.

Duan, H.; Hu, Q.; Wang, J.; Yang, L.; Xu, Z.; Liu, L.; Min, X.; Cai, C.; Ye, T.; Zhang, X.; and Zhai, G. 2025. FineVQ: Fine-Grained User Generated Content Video Quality Assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 3206–3217.

Duan, H.; Min, X.; Wu, S.; Shen, W.; and Zhai, G. 2024. UniProcessor: A Text-induced Unified Low-level Image Processor. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

GeminiTeam. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, 127: 398 – 414.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Hu, S.; Feng, M.; Nguyen, R. M. H.; and Lee, G. H. 2018. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ji, K.; Guo, Y.; Zhang, Z.; Zhu, X.; Tian, Y.; Liu, N.; and Zhai, G. 2025a. Medomni-45 {\deg}: A safety-performance benchmark for reasoning-oriented llms in medicine. *arXiv preprint arXiv:2508.16213*.

Ji, K.; Han, J.; Zhai, G.; and Liu, J. 2025b. Assessing the Capabilities of Generative Pretrained Transformer-4 in Addressing Open-Ended Inquiries of Oral Cancer. *International Dental Journal*, 75(1): 158–165.

Ji, K.; Wu, Z.; Han, J.; Jia, J.; Zhai, G.; and Liu, J. 2024. Application of 3D nnU-Net with residual encoder in the 2024 MICCAI head and neck tumor segmentation challenge. In *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*, 250–258. Springer.

Ji, K.; Wu, Z.; Han, J.; Zhai, G.; and Liu, J. 2025c. Evaluating ChatGPT-4's performance on oral and maxillofacial queries: Chain of Thought and standard method. *Frontiers in Oral Health*, 6: 1541976.

Jin, L.; and Jia, L. 2025. Embodied World Models Emerge from Navigational Task in Open-Ended Environments. arXiv:2504.11419.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. LLaVA-OneVision: Easy Visual Task Transfer. In *Proceedings of the Transactions on Machine Learning Research*.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024b. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. In *Proceedings of the International Conference on Learning Representations*.

Li, W.; Lai, Y.; Xu, L.; Xiangli, Y.; Yu, J.; He, C.; Xia, G.-S.; and Lin, D. 2023. OmniCity: Omnipotent City Understanding with Multi-level and Multi-view Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. In *Proceedings of the Conference on Neural Information Processing Systems*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; Chen, K.; and Lin, D. 2023b. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv:2307.06281*.

Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566.

Mai, J.; Chen, J.; Li, B.; Qian, G.; Elhoseiny, M.; and Ghanem, B. 2023. LLM as A Robotic Brain: Unifying Egocentric Memory and Control. arXiv:2304.09349.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. Accessed: 2025-01-23.

OpenAITeam. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Team, C. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. In *Anthropic*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution.

Wang, Z.; Xu, D.; Khan, R. M. S.; Lin, Y.; Fan, Z.; and Zhu, X. 2024b. LLMGeo: Benchmarking Large Language Models on Image Geolocation In-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Weyand, T.; Kostrikov, I.; and Philbin, J. 2016. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *Proceedings of the European Conference on Computer Vision*, 37–55. Springer International Publishing. ISBN 9783319464848.

Workman, S.; Souvenir, R.; and Jacobs, N. 2015. Wide-Area Image Geolocalization with Aerial Reference Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Yang, L.; Duan, H.; Tao, R.; Cheng, J.; Wu, S.; Li, Y.; Liu, J.; Min, X.; and Zhai, G. 2025a. ODI-Bench: Can MLLMs Understand Immersive Omnidirectional Environments? arXiv:2510.11549.

Yang, L.; Duan, H.; Zhu, Y.; Liu, X.; Liu, L.; Xu, Z.; Ma, G.; Min, X.; Zhai, G.; and Le Callet, P. 2025b. Omni2: Unifying Omnidirectional Image Generation and Editing in an Omni Model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 10103–10112. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Yeh, C.-H.; Wang, C.; Tong, S.; Cheng, T.-Y.; Wang, R.; Chu, T.; Zhai, Y.; Chen, Y.; Gao, S.; and Ma, Y. 2025. Seeing from Another Perspective: Evaluating Multi-View Understanding in MLLMs. arXiv:2504.15280.

Yue, X.; Ni, Y.; Zheng, T.; Zhang, K.; Liu, R.; Zhang, G.; Stevens, S.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Zhenzhu, Y.; Liu, Y.; Huang, W.; and Chen, W. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, Z.; Wang, J.; Guo, Y.; Wen, F.; Chen, Z.; Wang, H.; Li, W.; Sun, L.; Zhou, Y.; Zhang, J.; Yan, B.; Jia, Z.; Xiao, J.; Tian, Y.; Zhu, X.; Zhang, K.; Li, C.; Liu, X.; Min, X.; Jia, Q.; and Zhai, G. 2025. AIBench: Towards Trustworthy Evaluation Under The 45° Law. https://aiben.ch/.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; Zhou, W.; and Chen, Y. 2024. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517.

Zheng, Y.; Zhang, Z.; Min, X.; Duan, H.; and Zhai, G. 2025. LM Fight Arena: Benchmarking Large Multimodal Models via Game Competition. arXiv:2510.08928.

Zhu, C.; Wang, T.; Zhang, W.; Pang, J.; and Liu, X. 2024. LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; and et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.

Zhu, S.; Yang, T.; and Chen, C. 2021. VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhu, X.; and Yang, Y. 2024. CVGlobe: A Large-Scale Cross-View Geo-Localization Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.